2.5
2.8
2.2
3.2
3.6
2.0
4.0
100 mm
1.0
1.1
1.25
1.4
1.6
2.8
3.2
3.6
4.0
2.5
2.2
2.0
1.8
ABCDEFGHIJKLMNOPQRSTUVWXYZ
abcdefghijklmnopqrstuvwxyz
1234567890
1.0 mm
1.5 mm
2.0 mm
A5

DOCUMENT RESUME

ED 309 643 FL 018 115

AUTHOR Silva, Tony
TITLE A Review of the Research on the Evaluation of ESL Writing.
PUB DATE 89
NOTE 16p.; Paper presented at a Meeting of the Conference on College Composition and Communication (Seattle, WA, 1989).
PUB TYPE Information Analyses (070) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Correlation; *English (Second Language); Evaluation Criteria; *Evaluation Methods; Literature Reviews; Second Language Instruction; Test Construction; Writing (Composition); *Writing Evaluation

ABSTRACT

A review of the literature on the evaluation of writing in English as a Second Language (ESL) discusses research in four areas: (1) general discussions of basic issues in large-scale ESL composition evaluation; (2) accounts of the development of particular instruments and programs for measuring the ability of ESL writers; (3) reports of research looking for correlations between results from different ESL composition evaluation schemes; and (4) treatments of other related ESL writing assessment topics. A 46-item bibliography is included. (MSE)

CCCC 1989: SEATTLE, WASHINGTON

A REVIEW OF THE RESEARCH ON THE EVALUATION OF ESL WRITING

TONY SILVA, PURDUE UNIVERSITY

OVERVIEW

More and more, writing instructors and writing program administrators find themselves dealing with non-native-English-speaking students and facing the issue (among many others) of how these students' writing should be evaluated. While an abundance of information on the evaluation of first language writing is readily available, treatments of second language writing assessment are fewer and often less accessible--especially for those with little or no familiarity with the ESL literature. The purpose of this paper is to bring the existing second language composition evaluation literature to light by offering a classification and brief description of forty-seven relevant research reports done during the last fifteen years. The reports examined fall into four basic categories: (1) general discussions of basic issues in large scale ESL composition evaluation, (2) accounts of the development of particular instruments and programs for measuring the ability of ESL writers, (3) reports of research looking for correlations between results from different ESL composition evaluation schemes, and (4) treatments of other related ESL writing assessment topics.

GENERAL DISCUSSIONS OF BASIC ISSUES IN LARGE-SCALE ESL COMPOSITION EVALUATION

General discussions of issues in ESL composition evaluation are few and typically focus on the issue of the relative merits of direct (e.g. holistic, analytic and primary trait ratings) and indirect measures of writing ability (e.g., error-based scores, structural indices and multiple choice tests). Oller, in his Language Tests at School (1979), explores the questions of whether holistic ratings are as reliable as error-based scores. Low (1982) discusses the pros and cons of direct and indirect testing for an ESL audience and proposes a rationale for the construction, analysis, and evaluation of direct tests of L2 academic writing ability. Perkins (1983), in a very comprehensive article, reviews the nature, strengths, weaknesses, and evaluation of different types of ESL writing measures, including holistic, analytic, primary trait, and indirect writing measures and a number of standardized tests sometimes used to gauge ESL writing proficiency--like the TAS (Test of the Ability to Subordinate), TSWE (Test of Standard Written English), and MTELP (Michigan Test of English Language Proficiency). Bridgeman and Carlson (1983) and Carlson et al (1985), while focusing mainly on a more particular topic (the development of the TWE [Test of Written English]), provide brief but useful overviews of a number of issues in ESL composition evaluation--including functionally-based communicative competencies, field-specific writing task demands, and perspectives from contrastive rhetoric--in their introductory/background sections.

THE DEVELOPMENT OF ESL COMPOSITION EVALUATION PROGRAMS

Direct Measurement: Holistic

Published accounts of work on holistic evaluation programs for ESL composition come basically from two sources: American institutions of higher

education and from ETS (Educational Testing Service). Informative reports on development, use, and evaluation of holistic instruments have been done by Reid and O'Brien (1981) at Colorado State University, by Robinson (1982) at St. Edward's University (Texas), and by Carr (1983) at the University of San Francisco.

The most extensive work on holistic measurement of ESL writing, by far, has been done by researchers working under the auspices of ETS on the TWE. Bridgeman and Carlson (1983)--summarized in Bridgeman and Carlson (1984)--surveyed faculty from thirty-four universities (six disciplines; 190 departments) to ascertain the academic writing skills needed by graduate and undergraduate international students in the USA. Their major findings included the following: (1) faculty saw writing skills as even more important for international students after graduation than while at school; (2) all disciplines required some writing from first-year students; (3) departments varied with regard to the particular writing skills they viewed as most important; (4) faculty said they relied more on discourse-level than on sentence-level features in evaluating student writing; (5) discourse-level writing skills for L1 and L2 writers were perceived as similar; differences were perceived at the levels of sentence and word and in overall writing ability; (6) departments varied in their preference for topic types; and (7) disciplines do not agree on writing task demands or mode of discourse best suited for evaluating entering students.

Carlson et al (1985), a follow-up on Bridgeman and Carlson (1983), describes the field testing of topics developed in the earlier study and reports on the correlation of the results (holistic scores) with scores on the TOEFL (Test of English as a Foreign Language), the GRE (Graduate Record Exam), and a multiple choice writing test. The major findings here include: (1) there was a close relationship between holistic, analytic, and objective scores;

therefore, it was felt that holistic scores alone would be sufficient for large-scale evaluation purposes; (2) high correlations within and across selected topic types were found: (3) the scores of raters with backgrounds in ESL, English, and other disciplines all correlated highly; (4) high correlations were found between holistic ratings and TOEFL scores--but each was judged to reliably measure an aspect of writing not assessed by the other. Carlson and Bridgeman (1986) summarizes all the research on the TWE conducted by ETS, i.e., Bridgeman and Carlson (1983) plus Carlson et al (1985).

Additional discussion of the TWE project has come both from sources inside as well as outside ETS. Stansfield (1986), speaking for ETS, provides a history of the TWE, addressing it in terms of motivation and research and development. This account is a technical treatment of the project aimed at measurement researchers. Stansfield and Webster (1986), also on ETS' behalf, provide a brief description of the TWE--covering its motivation, history, topics, scoring, and the use of results from it. This is a popular treatment aimed at test-score users--ESL teachers and administrators. Finally, Greenberg (1986) presents a review of TOEFL Research Reports 15 and 19, i.e., Bridgeman and Carlson (1983) and Carlson et al (1985). The review is generally favorable, but Greenberg expresses r ·ervations about: (1) topic complexity (given a thirty-minute test format); (2) a bias in favor of the reactions of graduate faculty; (3) equivalence of topic types used; (4) the effects of time constraints on prewriting and revising; and (5) the use of a single writing sample as a basis for evaluation.

Direct Measurement: Analytic

Jacobs et al (1981) is, by far, the most comprehensive treatment of ESL composition evaluation that exists at present. This text, Testing ESL Composition, is a handbook designed to guide ESL practitioners in planning, conducting, evaluating and using the results of direct pragmatic tests of composition, particularly an analytic instrument--the ESLCP (ESL Composition Profile)--which renders a total score (on a hundred-point scale) and part scores for content, organization, language use, vocabulary, and mechanics. The handbook's contents include: (1) background information on composition testing; (2) guidelines for developing an ESL composition testing program (particularly, discussions of establishing the purpose of evaluation, preparing the writing task, planning evaluation procedures, selecting and training raters, administering the test, evaluating the compositions, interpreting test scores, and making decisions and recommendations from test scores); (3) a reader guide for composition evaluation (focusing on using the ESLCP and ensuring efficiency and reliability; sample essays with ratings are also supplied); and (4) descriptive statistics and other information on the use of the ESLCP at the University of Texas at Austin.

Hamp-Lyons (1986) discusses an analytic scoring procedure for discipline-specific content-focused writing tasks. In particular, she describes the development of four, nine-point scales (assessing content, format, linguistic features, and task fulfillment) designed for use with the writing subtest of the British Council's English Language Testing Service (ELTS) English proficiency test. Henning & Davidson (1987) do a scalar analysis of an analytic rating scheme used at UCLA. They focus on subscale difficulties, weights, and intervals; reliability estimates; fit validity; and the nature of misfitting performances. They conclude that though the instrument is highly accurate, distinguishing ratings at the mid points of scales is difficult.

Other smaller-scale analytic scoring programs are also reported in the literature. Aghbar (1983) reports on the adaption and use of an analytic instrument, assessing rhetorical structure, grammar, vocabulary, and spelling, at George Mason University. Zughoul and Kambal (1983) discuss the development, use, and evaluation of analytic scales, measuring structure, content, organization, and mechanics for each of three proficiency levels (beginning, intermediate, and advanced), at the University of Texas at Austin and a Yarmouk University in Jordan. Brown and Bailey (1984) describe an analytic scoring instrument designed for use with upper-intermediate ESL university students at the University of California, Los Angeles. The instrument involves five equally-weighted criteria--organization, logical development of ideas, grammar, mechanics, and style.

Indirect Measurement

Various indirect measures of ESL students' writing ability, designed for large-scale evaluation, have been described in the literature. Gipps and Ewen (1974) present a scoring system employing mean T-unit length and the sum of intelligibility ratings of T-units (0=unintelligible; 1=partially intelligible; 2=completely intelligible; 3=completely accurate), i.e., complexity and intelligibility scores for compositions. A series of publications by D. Davidson involve an indirect measure that he devised--the TAS (Test of the Ability to Subordinate). In his dissertation (1976a)--summarized in D. Davidson (1976b)--he describes the development and testing of the TAS: a test that requires students to combine sentences and that focuses on nine grammatical structures commonly believed to cause problems for inexperienced ESL writers. Davidson reports here that the TAS was found to be reliable and to correlate highly with results from

the MTELP and with holistic writing sample scores (.74). D. Davidson (1978a) and (1978b) are a standardized published version of the TAS and a TAS user's manual, respectively.

Brodkey and Young (1981) discuss an indirect measure called a "composition correctness score." The procedure for computing this score involves (1) finding all errors in the first 250 words of a writing sample, (2) assigning a weight to each error (3=severe; 2=moderate; 1=minor); and (3) dividing 250 (the number of words analyzed) by the sum of the weighted errors--the quotient is the composition correctness score. The authors report that these scores correlated highly with holistic scores and discriminated among four narrow-range proficiency levels. Finally, J. Davidson (1985a, 1985b) describes a multiple choice ESL writing placement test. The test consists of forty items which focus on eight "rhetorical concerns:" (1) text-diagram/chart correlation; (2) sentence deletion in spatially, chronologically, and heuristically structured texts; (3) topicalization; (4) cohesion; (5) topic sentence recognition; (6) coherence; (7) outlining; and (8) subordination. The test was reported to be reliable, but not to correlate well with holistic or TOEFL scores.

RELATIONSHIPS BETWEEN ESL COMPOSITON MEASURES

Correlation Studies: Indirect Measures and Holistic Scores

A number of studies focus primarily on the correlation of particular indirect measures of ESL writing ability and holistic scores for writing samples. Anderson (1980), looking primarily at cohesion, found that holistic scores correlated (1) insignificantly with frequency of cohesion; (2) negatively with frequency of reference cohesion; and (3) positively with frequency of

conjunction usage and frequency of correct conjunction usage. Arthur (1980) found significant correlations between holistic ratings and length, frequency of grammatical errors, and frequency of spelling mistakes. Homburg (1980) discovered a number of relationships between particular syntactic maturity measures and holistic scores; Perkins and Homburg (1980) found that number of errors per sample and per T-unit correlated significantly with holistic ratings.

Perkins (1980) reported that the variables of number of error-free T-units, number of words in error-free T-units, errors per T-unit, and total errors correlated highly with holistic scores while TSWE scores did not. Perkins found no significant correlation between holistic ratings and TAS scores in his 1981 study; however, he reported a substantial correlation between these two in his 1984 paper. Kameen (1983) discussed three variables that distinguished good and poor ESL writers (i.e., those with high and low holistic scores). These were T-unit length, clause length, and incidence of passive voice. Finally, Homburg (1984) found that five factors accounted for 84% of the variance in holistic scores in his sample. These factors included (1) second-degree errors per T-unit, (2) dependent clauses per composition, (3) words per sentence, (4) coordinating conjunctions per composition, and (5) error-free T-units per composition.

Correlation Studies: Other Two-Way Relationships

At least five studies report on two-way relationships that do not juxtapose holistic and indirect measure scores. Chance (1973) discovered high correlations between scores on a multiple-choice composition test with writing course grades and MTELP subtest scores. Perkins (1982) reported that results from holistic and analytic evaluations correlated highly, that they yielded the same results and

measured the same constructs. Mullen (1977, 1980--these are two versions of the same study) compared scores from holistic and analytic (structure, organization, vocabulary, and quantity) measures and found that (1) all four part scales together did a better job in predicting holistic ratings than any one, two or three; (2) vocabulary ratings correlated most highly with holistic scores; and (3) organization scores evidenced the weakest relationship with holistic ratings. Lim (1983) reported high correlations between L2 proficiency and number of error-free T-units, words per error-free T-unit, and words per T-unit. Last, Perkins (1984) found no concurrent validity between ESLCP ratings and scores on the TAS or RET (Revising and Editing Test), two standardized indirect writing tests.

Correlation Studies: Relationships between Three or More Variables

A couple of studies reported on correlations across three or more variable categories. Flahive and Snow (1980) found high correlations between (1) placement level and T-unit length and clause/T-unit ratio, and (2) holistic scores and clause/T-unit ratio (at low proficiency levels) and T-unit length (at high proficiency levels). And in her 1980 study, Kaczmarek reported high correlations among all four of her variables: holistic ratings, analytic ratings, error-based scores, and scores on a multiple-choice indirect writing test.

OTHER ISSUES IN ESL COMPOSITION EVALUATION

Other evaluation-related issues discussed include rater judgements writing topics, and research agendas. Carney (1973) addressed the characteristics of experienced and inexperienced raters of ESL writing. She reported that while

experienced raters stressed organization, meaning, and communication and looked for specific elements in papers, inexperienced judges stressed mechanics and formed broad impressions of student texts. It was also noted that experienced raters made more uniform judgements; differed from the experienced raters in the weights assigned to different criteria; and observed a hierarchy of features--looking first at content and organization, second at rhetorical devices, and third at errors. In investigating the judgement of raters using an analytic scale on ESL compositions, Mullen (1977, 1980) determined that some rater pairs were reliable and equivalent (assigned similar scores); some pairs were reliable but not equivalent; and some pairs were neither reliable nor equivalent.

In 1985, on the basis of data from a large-scale university-level composition testing program, McDaniel reported that raters judged the compositions of ESL and native-English-speaking writers differently. Robinson (1985) found, in a controlled study, that (1) English teachers (teachers of English to native speakers of English) rated ESL compositions significantly lower than did ESL teachers, and (2) ESL compositions rewritten by native speakers (changed only in terms of handwriting and in the correction of grammar and spelling errors) were rated significantly higher than the originals by both English and ESL teachers. Leonhardt (1985), investigating the effect of assigned versus open topics on ESL students writing scores, found that when level of L2 proficiency was controlled for, there was no significant correlation between topic type and analytic rating (ESLCP score).

Finally, Perkins (1986) suggests a long-term research agenda for ESL composition evaluation, recommending further rigorous investigation of (1) possible types of rating analyses (holistic, analytic), (2) various methods of estimating the reliability of composition ratings, (3) effects of different

purposes of writing on samples elicited during the testing process, (4) the measurement properties of rating scales, (5) estimation of rater effects, (6) the number and nature of constraints underlying writing ability, (7) the assessment of errors, (8) the role of writing apprehension in composition assessment, and (9) the need for systematic, replicable analyses of the composing process.

CONCLUSION

In conclusion, it is not suggested that this literature review is complete or exhaustive; however, I think it does provide a reasonably representative sample of work in large-scale ESL composition assessment done to date. And while it is certainly a modest collection, one which has a lot of obvious gaps and whose studies are quite uneven in terms of breadth, depth, quality, and significance, I believe it still represents a substantial and important body of information and a valuable resource for instructors and administrators who need to evaluate the work of ESL writers.

BIBLIOGRAPHY

Aghbar, A. (1983). Guided impressionistic scoring of ESL compositions. ERIC Document 242195.

Anderson, P. (1980). Cohesion as an index for written composition of ESL learners. ERIC Document 198529.

Arthur, B. (1980). Short-term changes in EFL composition skills. In C. Yorio, K. Perkins, and J. Schacter (Eds.), On TESOL '79: The learner in focus. Washington, DC: TESOL.

Bridgeman, B. (1986). Testing ESL student writers. In K. Greenberg, H. Wiener, & R. Donovan (Eds.), Writing assessment: Issues and strategies. New York: Longman

Bridgeman, B. & Carlson, S. (1983). A survey of academic writing tasks required of graduate and undergraduate foreign students. TOEFL Research Report 15. Princeton, NJ: Educational Testing Service.

Bridgeman, B. & Carlson, S. (1984). Survey of academic writing tasks. Written Communication, 1(2), 247-280.

Brodkey, D. & Young, R. (1981). Composition correctness scores. TESOL Quarterly, 15, 403-411.

Brown, J. & Bailey, K. (1984). A categorical instrument for scoring second language writing skills. Language Learning, 34(4), 21-42.

Carlson, S., Bridgeman, B., Camp, R., & Waanders, J. (1985). Relationship of admission test scores to writing performance of native and non-native speakers of English. TOEFL Research Report 19. Princeton, NJ: Educational Testing Service.

Carney, H.(1973). An inquiry into criteria for composition evaluation in English as a foreign language. Dissertation Abstracts International, 34(8), 5139-A.

Carr, M. (1983). A five-step evaluation of a holistic essay-evaluation process. ERIC Document 238263.

Chance, L. (1973). The development of an objective composition test for non-native speakers of English. Dissertation Abstracts International, 34(12), 7511-A.

Davidson, D. (1976). Assessing writing ability of ESL college freshman. ERIC Document 135247.

Davidson, D. (1976). Development and validation of a diagnostic examination of ability to subordinate in writing for ESL college students. Dissertation

Abstracts International, 37(9), 5790-A.

Davidson, D. (1978a). Teacher's manual for test of test of ability to subordinate. New York: Language Innovations, Inc.

Davidson, D. (1978b). Test of ability to subordinate. New York: Language Innovations, Inc.

Davidson, J. (1985a). An indirect measure of writing skills for student placement in freshman and pre-freshman ESL courses: Part I. TECFORS, 8(3), 1-7.

Davidson, J. (1985b). An indirect measure of writing skills for student placement in freshman and pre-freshman ESL courses: Part II. TECFORS, 8(4), 7-9.

Flahive, D. & Snow, B. (1980). Measures of syntactic complexity and evaluating ESL compositions. In J. Oller & K. Perkins (Eds.), Research in language testing. Rowley, MA: Newbury House.

Gipps, C. & Ewen, E. (1974). Scoring written work in English as a second language: The use of the T-unit. Educational Research, 16, 121-125.

Greenberg, K. (1986). The development and validation of the TOEFL writing test: A discussion of TOEFL Research Reports 15 and 19. TESOL Quarterly, 20(3), 531-544.

Hamp-Lyons, L. (1986). Testing writing across the curriculum. Papers in Applied Linguistics-Michigan, 2(1), 16-29.

Henning, G., & Davidson, F. (1987). Scalar analysis of composition ratings. ERIC Document 287285.

Homburg, T. (1980). A syntactic complexity measure of attained ESL writing proficiency. Unpublished Master's Thesis, Southern Illinois University, Carbondale.

Homburg, T. (1984). Holistic evaluation of ESL compositions: Can it be evaluated

objectively? TESOL Quarterly, 18, 87-107.

Jacobs, H., Zinkgraf, S., Wormuth, D., Hartfiel, V. & Hughey, J. (1981). Testing ESL composition: A practical approach. Rowley, MA: Newbury House.

Kaczmarek, C. (1980). Scoring and rating essay tasks. In J. Oller & K. Perkins (Eds.), Research in Language Testing. Rowley, MA: Newbury House.

Lim, H. (1983). Using T-unit measures to assess writing proficiency of university ESL students. RELC Journal, 14(2), 35-43.

McDaniel, B. (1985). Ratings versus equity in the evaluation of writing. ERIC Document 260459.

Mullen, K. (1977). Using rater judgements in the evaluation of writing proficiency for non-native spekers of English. In H. Brown, C. Yorio, & R. Crymes (Eds.), On TESOL '77: Teaching and learning ESL--Trends in research and practice. Washington, DC: TESOL.

Mullen, K. (1980). Evaluating writing proficiency. In J. Oller and K. Perkins (Eds.), Research in language testing. Rowley, MA: Newbury House

Oller, J. (1979). Language tests at school. London: Longman.

Perkins, K. (1980). Using objective methods of attained writing proficiency to discriminate among holistic evaluations. TESOL Quarterly, 14(1), 61-69.

Perkins, K. (1981). The test of the ability to subordinate: Predictive and concurrent validity for attained ESL composition. ERIC Document 217734.

Perkins, K. (1982). An analysis of the robustness of composition scoring schemes. ERIC Document 217723.

Perkins, K. .1983). On the use of composition scoring techniques, objective measures, and objective tests to evaluate ESL writing ability. TESOL Quarterly, 17, 651-671.

Perkins, K. (1984). A regression analysis of direct and indirect measures of English as a second language writing compositions. ERIC Document 275170.

Perkins, K. (1986). A proposed research program for ESL composition evaluation. ERIC Document 290155

Perkins, K. & Homburg, T. (1980). Three different statistical analyses of objective measures of attained ESL writing proficiency. In R. Silverstein (Ed.), Proceedings of the third annual conference on frontiers in language proficiency and dominance testing. Carbondale, IL: Southern Illinois University.

Perkins, K. & Parish, C. (1984). Direct versus indirect measures of writing proficiency: Research in ESL composition. ERIC Document 243306.

Reid, J. & O'Brien, M. (1981). The application of holistic grading in an ESL writing program. ERIC Document 221044.

Robinson, T. (1982). Holistic scoring of essays. TECFORS, 5(2), 7-9.

Robinson, T. (1985). Evaluating foreign students' compositions: The effects of rater background and of handwriting, spelling, and grammar. Dissertation Abstracts International, 45(3), 2951-A.

Stansfield, C. (1986). A history of the Test of Written English. Language Testing, 3(2), 224-234.

Stansfield, C. & Webster, R. (1986). The new TOEFL writing test. TESOL Newsletter, 20(5), 17-18.

Zughoul, M. & Kambal, M. (1983). Objective evaluation of ESL composition. IRAL, 21, 87-103.